

6. Regressionsanalyse, mindste kvadraters metode og bedste rette linje

Data, som vi indsamler for at opnå viden om omverden, er i mange situationer behæftet med variation, eller støj, som vi også vil kalde det. Eller måske er der tale om, at der er skjulte variable på spil, hvis indflydelse ikke fanges af den matematiske model, vi lige nu anser som den bedste? Derfor er der i matematik også udviklet metoder til at "trænge igennem støjen" og nå ind til de underliggende sammenhænge. Metoderne går under den fælles betegnelse *regressionsanalyse*, og det grundlæggende værktøj, vi anvender, er den såkaldte *mindste kvadraters metode*.

Lad os introducere metoden med et praktisk eksempel:

6.1 Eksempel: Bedste rette linje gennem et sæt af datapunkter

(Dette afsnit er lånt fra HEM2, projekt 8.5)

I 1880'erne indsamlede Francis Galton (1822-1911) et datasæt, der indeholdt sammenhørende målinger af fædres højder og deres førstefødte sønners højder som voksne. Der er målinger for i alt 952 par af fædre og sønner. Formålet med arbejdet var dengang at få en bedre indsigt i, hvordan egenskaber nedarves. Historisk set var det analysen af dette datasæt, hvor du til højre ser de første 10 målinger (angivet i cm), der gav anledning til det umiddelbart besynderlige navn "regressionsanalyse". Regression er det modsatte af progression og betyder tilbageskridt. Konklusionen på undersøgelsen var nemlig, at særligt høje fædre i gennemsnit fik knap så høje sønner – og det opfattede Galton som et tilbageskridt! Men samtidig fik små fædre sønner, der også var små, men dog lidt højere.	nummer	fars højde	sønnens højde
	1	186,9	183,4
	2	184,6	172,0
	3	185,0	179,0
	4	182,1	165,4
	5	179,4	155,4
	6	178,4	160,3
	7	179,7	165,0
	8	179,7	168,7
	9	176,5	160,3
	10	173,4	157,5
De første 10 ud af Galtons datasæt på i alt 952			

Øvelse 6.1 Udfør lineær regression ved hjælp af et matematisk værktøjsprogram

a) Udfør et punktplot af datasættet her, og kommenter, om plottet giver anledning til at tro på, at der findes en bagvedliggende lineær sammenhæng.

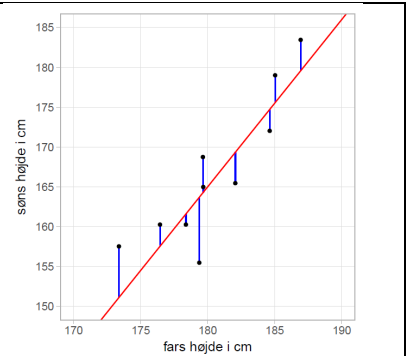
b) Udfør lineær regression på datasættet. Du skal få en linje nogenlunde som den røde linje nedenfor. Samtidig giver programmet dig formlen: $y = 1.94x - 183.58$. Hvor kommer den fra?

Fremgangsmåden er følgende:

1. Vi måler **den lodrette afstand** mellem linjen og datapunkterne – det er længden af de blå linjestykker. Kaldes vi fædrenes højder for $x_1, x_2, \dots, x_{10}$ og sønnernes højder for $y_1, y_2, \dots, y_{10}$, og er regressionslinjens forskrift $y = a \cdot x + b$, så er den lodrette afstand for det første datapunkt:

$$y_1 - (a \cdot x_1 + b) \quad (*)$$

Overvej selv dette!



2. Vi ønsker ét samlet tal, der skal måle afstanden mellem linjen og datasættet. Læg mærke til, at nogle gange, fx ved punkt nr 3 og 4, vil formlen (*) give et negativt tal. Men hver afstand skal give et positivt bidrag. Derfor vælger vi at kvadrere disse tal og bagefter summere. (I videregående statistik argumenteres for denne fremgangsmåde).

3. Det samlede udtryk for afstanden mellem linjen og datasættet er derfor:

$$S = (y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + \dots + (y_{10} - (a \cdot x_{10} + b))^2$$

4. Udtrykket i punkt 3 kan skrives op med *ethvert* lineært udtryk, $y = a \cdot x + b$. Blandt alle disse uendeligt mange udtryk for S , vælger vi nu den linje, der giver *det mindste tal*. Det var det samme Oscar og Casper søgte efter med øjemål i øvelsen på tavlen. Når vi skal være mere præcise, kommer differentialregningen på banen – det sker i afsnit 6.3

Bemærk: Der ligger i det foregående nogle subjektive valg, der ikke kan begrundes alene ud fra den matematik vi har kendskab til.

- For det første vælger vi at måle afstanden lodret. Men den mindste afstand fra et punkt til en linje findes som den *vinkelrette* afstand. Dette giver imidlertid nogle ganske klodsede udtryk, som det er svært at regne på. Endvidere er det svært at generalisere metoden til fx polynomisk regression. Så vi *vælger* at se på den lodrette afstand. Det giver ikke alene en fordel mht udregninger, men betyder også, at vi ser på forskellen mellem den empiriske dataværdi og modelværdien.

- For det andet vælger vi at se på kvadratet på afstanden og ikke kun afstanden. Det er også et valg, der bl.a. er bestemt af, at afstand måles som den *numeriske* værdi, og denne størrelse er svær at håndtere i store udtryk. Men det betyder, at *afvigere* kan komme til at betyde uforholdsmæssigt meget, da afstandene kvadreres.

I den gren af matematik, der hedder *Lineær algebra* opdager man, at den samlede sum af afstandskvadrater mellem et datasæt og en linje er et udtryk for en afstand i et n -dimensionalt rum, nemlig *afstanden fra det punkt, som datasættet repræsenterer til linjen*. Hvorfor kvadrater? Fordi afstande også i n dimensioner findes vha. Pythagoras sætning! Du kan læse mere herom i HEM3 kapitel 9.

Definition: Regressionslinje

Den linje, der passer bedst muligt til givne dataværdier, kaldes *regressionslinjen* (af og til *tendenslinjen*), og vi siger, at linjen er fremkommet ved at lave *lineær regression*. Bedst muligt er bestemt ved *mindste kvadraters metode*, se nedenfor.

6.2 Hvor god er tilnærmelsen - residualplot

Selv om regressionslinjen passer nogenlunde til punkterne, så kan vi ikke vide med sikkerhed, at der også er en egentlig årsagssammenhæng. Ofte skal andre fag bidrage til at

afgøre, om der faktisk er tale om en årsagssammenhæng og ikke kun en matematisk sammenhæng.

Et værktøj til at svare på, hvor godt modelværdierne passer med måledata, er det såkaldte *residualplot*. Et residualplot giver et grafisk billede af forskellen mellem *dataværdierne* og de beregnede *modelværdier*, dvs længden (regnet med fortegn) af de blå linjestykker. Vi kan derved få et visuelt indtryk af, om forskellen mellem model og virkelighed kan tilskrives tilfældigheder eller synes at være systematiske og dermed udtryk for, at der er nogle sammenhænge, vi ikke har styr på.

Øvelse 6.2: Residualer og residualplottet på datasættet

- Opstil selv en tabel over residualerne, dvs. forskellen mellem *dataværdierne* og de beregnede *modelværdier*.
- Undersøg, hvordan dit værktøjsprogram kan gøre det pr automatik.
- Tegn residualplottet. Kommentér



6.3. Formlen for lineær regression

S , der er skrevet op i afsnit 6.1, og som måler kvadratsummen af afvigelserne, kan betragtes som en funktion af de to variable a og b , $S(a,b)$, og de værdier af a og b , der giver den mindste kvadratsum, findes som et minimumssted for funktionen $S(a,b)$. Funktioner af to variable behandles med principielt samme metode, som vi kender fra funktioner af én variabel:

- Vi differentierer først mht b og løser ligningen: $S'_b(a,b)=0$
- Vi differentierer dernæst mht. a og løser ligningen $S'_a(a,b)=0$

Resultatet af det hele kan sammenfattes i denne sætning, som vi beviser nedenfor

Sætning 7: Formlen for den lineære regressionslinje

Lad der være givet et datasæt bestående af n punkter, (x_i, y_i) , hvor $i=1..n$.

Vi indfører betegnelserne $\bar{x}$ og $\bar{y}$ for middelværdierne (gennemsnittene):

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^n x_i, \quad \text{og} \quad \bar{y} = \frac{1}{n} \cdot (y_1 + y_2 + \dots + y_n) = \frac{1}{n} \cdot \sum_{i=1}^n y_i$$

Koefficienterne i regressionslinjen (*den bedste rette linje*) betegnes $\hat{a}$ og $\hat{b}$ og beregnes således:

$$\hat{a} = \frac{\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$$

Bevis.

Las os nu sige, vi har n datapunkter. Vi opskriver udtrykket for S :

$$S = (y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + \dots + (y_n - (a \cdot x_n + b))^2 = \sum_{k=1}^n (y_k - (a \cdot x_k + b))^2$$

Sumtegnet er en kompakt og overskuelig måde at skrive lange udtryk på. Men man skal lige vænne sig til det, så vi vil veksle.

Først vil vi gange parenteserne ud. Den generelle parentes udregnes:

$$(y_k - (a \cdot x_k + b))^2$$

Benyt

kvadratsætningen på de to led:

$$= y_k^2 + (a \cdot x_k + b)^2 - 2 \cdot y_k \cdot (a \cdot x_k + b)$$

Benyt kvadratsætningen på

$$(a \cdot x_k + b)^2 :$$

$$= y_k^2 + ((a \cdot x_k)^2 + b^2 + 2 \cdot a \cdot x_k \cdot b) - 2 \cdot y_k \cdot (a \cdot x_k + b)$$

Hæv de første og gange ind i den sidste

parentes

$$= y_k^2 + a^2 \cdot x_k^2 + b^2 + 2 \cdot a \cdot x_k \cdot b - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot b \quad (*)$$

Vi skal bestemme a og b så dette udtryk minimeres. Læg mærke til, at x 'erne og y 'erne er konstanter. Det er jo de givne datapunkter. Det betyder, at udtrykket kan ses som en funktion af to variable, a , og b . Grafen for funktionen $S(a,b)$ er en flade der ligger over (a,b) -planen, hvor a og b er de uafhængige variable afsat ud af henholdsvis 1. og 2. aksens. Vi leder efter et minimum for denne flade.

Men det er klart, at findes et sådant minimum, hvor hele fladen er "nede i en bølgedal", så er dette minimumspunkt også minimum for den funktion af a , hvis graf vi får ved at lægge et lodret snit gennem minimumspunktet og parallel med a -aksen. Og tilsvarende er det minimumspunkt for den funktion af b hvis graf vi får ved at lægge et lodret snit gennem minimumspunktet og parallel med b -aksen.

Derfor vil vi nu bestemme først b ved at holde a fast (dvs vi ser på snitkurven parallel med b -aksen). Og derefter a efter samme opskrift. (*Gjorde vi det i den modsatte rækkefølge, ville udregningerne blive meget mere komplicerede!*)

Vi differentierer udtrykket (*) mht b :

$$\frac{d}{db} (y_k^2 + a^2 \cdot x_k^2 + b^2 + 2 \cdot a \cdot x_k \cdot b - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot b)$$

$$= 2 \cdot b + 2 \cdot a \cdot x_k - 2 \cdot y_k$$

Det var det k'te led, og nu summerer vi og sætter lig med 0 – overvej nøje, at vi kan gøre det, ved fx at skrive det ud som en stor sum:

$$\sum_{k=1}^n (2 \cdot b + 2 \cdot a \cdot x_k - 2 \cdot y_k) = 0$$

Vi summerer de tre led hver for sig, og rykker

det sidste over:

$$\sum_{k=1}^n 2 \cdot b + \sum_{k=1}^n 2 \cdot a \cdot x_k = \sum_{k=1}^n 2 \cdot y_k$$

Vi forkorter med 2:

$$\sum_{k=1}^n b + \sum_{k=1}^n a \cdot x_k = \sum_{k=1}^n y_k$$

Vi sætter a udenfor sumtegn, dvs parentes, og

tæller antal b :

$$n \cdot b + a \cdot \sum_{k=1}^n x_k = \sum_{k=1}^n y_k$$

Vi dividerer med: n

$$b + a \cdot \frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{n} \sum_{k=1}^n y_k$$

Vi anvender definitionen på gennemsnit, som kaldes $\bar{x}$ og $\bar{y}$

:

$$b + a \cdot \bar{x} = \bar{y}$$

Vi isolerer b :

$$b = \bar{y} - a \cdot \bar{x}$$

Det er formelen for b i sætningen.

Derefter vil vi nu bestemme a ved at holde b fast (dvs vi ser på snitkurven parallel med a -aksen):

Men først indsættes den b -værdi, som vi nu kender i udtrykket (*)

$$S(a) = y_k^2 + a^2 \cdot x_k^2 + (\bar{y} - a \cdot \bar{x})^2 + 2 \cdot a \cdot x_k \cdot (\bar{y} - a \cdot \bar{x}) - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot (\bar{y} - a \cdot \bar{x})$$

Gang parenteserne ud:

$$S(a) = y_k^2 + a^2 \cdot x_k^2 + \bar{y}^2 + a^2 \cdot \bar{x}^2 - 2 \cdot \bar{y} \cdot a \cdot \bar{x} + 2 \cdot a \cdot x_k \cdot \bar{y} - 2 \cdot a^2 \cdot x_k \cdot \bar{x} - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot \bar{y} + 2 \cdot y_k \cdot a \cdot \bar{x}$$

Dette udtryk differentieres:

$$\frac{d}{da} (y_k^2 + a^2 \cdot x_k^2 + \bar{y}^2 + a^2 \cdot \bar{x}^2 - 2 \cdot \bar{y} \cdot a \cdot \bar{x} + 2 \cdot a \cdot x_k \cdot \bar{y} - 2 \cdot a^2 \cdot x_k \cdot \bar{x} - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot \bar{y} + 2 \cdot y_k \cdot a \cdot \bar{x})$$

$$= 2 \cdot a \cdot x_k^2 + 2 \cdot a \cdot \bar{x}^2 - 2 \cdot \bar{y} \cdot \bar{x} + 2 \cdot x_k \cdot \bar{y} - 4 \cdot a \cdot x_k \cdot \bar{x} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \bar{x}$$

Saml a -leddene

$$= 2 \cdot a \cdot (x_k^2 + \bar{x}^2 - 2 \cdot x_k \cdot \bar{x}) - 2 \cdot \bar{y} \cdot \bar{x} + 2 \cdot x_k \cdot \bar{y} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \bar{x}$$

Anvend en kvadratsætning:

$$= 2 \cdot a \cdot (x_k - \bar{x})^2 - 2 \cdot \bar{y} \cdot \bar{x} + 2 \cdot x_k \cdot \bar{y} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \bar{x}$$

Sæt henh. $-2 \cdot \bar{y}$ og $+2 \cdot y_k$ uden for parentes:

$$= 2 \cdot a \cdot (x_k - \bar{x})^2 - 2 \cdot \bar{y} \cdot (\bar{x} - x_k) + 2 \cdot y_k \cdot (\bar{x} - x_k)$$

Kontroller! – og sæt nu $(\bar{x} - x_k)$ udenfor

$$= 2 \cdot a \cdot (x_k - \bar{x})^2 - 2 \cdot (\bar{x} - x_k) \cdot (\bar{y} - y_k)$$

Det var det k'te led, og nu summerer vi og sætter lig med 0.

$$\sum_{k=1}^n (2 \cdot a \cdot (x_k - \bar{x})^2 - 2 \cdot (\bar{x} - x_k) \cdot (\bar{y} - y_k)) = 0$$

Summen splittes op og vi rykker over:

$$\sum_{k=1}^n 2 \cdot a \cdot (x_k - \bar{x})^2 = \sum_{k=1}^n 2 \cdot (\bar{x} - x_k) \cdot (\bar{y} - y_k)$$

Forkort med 2 og sæt a uden for:

$$a \cdot \sum_{k=1}^n (x_k - \bar{x})^2 = \sum_{k=1}^n (\bar{x} - x_k) \cdot (\bar{y} - y_k)$$

Isoler a :

$$a = \frac{\sum_{k=1}^n (\bar{x} - x_k) \cdot (\bar{y} - y_k)}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

Det er formelen for a i sætningen

Øvelse 6.3 Udfør lineær regression ved at beregne koefficienterne

Anvend sætning 2 til at beregne formelen for den lineære regression på de 10 datapunkter, og sammenlign med resultatet i øvelse 6.1. Det er lettest at opstille beregningen i et regneark, hvor dataværdierne er skrevet ind i to kolonner.

Øvelse 6.4 Undersøg et datasæt med regression og residualplot

Du kan [her \(link til 13\)](#) finde et andet lille datasæt, som du skal undersøge ved at udføre lineær regression og tegne et residualplot.